

#1358-101

A

68188 U.S. PTO  
05/23/97

FISH RICHARDSON P.C.

Frederick P. Fish  
1855-1930  
W.K. Richardson  
1859-1951

4225 Executive Square  
Suite 1400  
La Jolla, California  
92037

Telephone  
619 678-5070

Facsimile  
619 678-5099

May 23, 1997

Attorney Docket No.: 07300/034001

Commissioner of Patents and Trademarks  
Washington, DC 20231

Presented for filing is a new original patent application of:

Applicant: JEFFREY SKOLNICK, MARIUSZ MILIK, AND ANDRZEJ KOLINSKI  
Title: PREDICTION OF RELATIVE BINDING MOTIFS OF BIOLOGICALLY ACTIVE PEPTIDES AND PEPTIDE MIMETICS

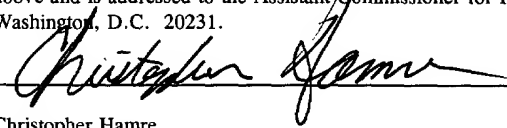
Enclosed are the following papers, including all those required for a filing date under 37 CFR §1.53(b):

Pages of Specification	15
Pages of Claims	6
Pages of Abstract	1
Pages of Declaration	[To Be Filed At A Later Date]
Sheets of Drawing	3

Basic filing fee	\$ 770.00
Total claims in excess of 20 times \$22.00	88.00
Independent claims in excess of 3 times \$80.00	240.00
Multiple dependent claims	260.00
Total filing fee:	\$ 1358.00

"EXPRESS MAIL" Mailing Label Number EM122760615US  
Date of Deposit May 23, 1997

I hereby certify under 37 CFR 1.10 that this correspondence is being deposited with the United States Postal Service as "Express Mail Post Office To Addressee" with sufficient postage on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

  
Christopher Hamre

BOSTON  
HOUSTON  
NEW YORK  
SOUTHERN CALIFORNIA  
SILICON VALLEY  
TWIN CITIES  
WASHINGTON, DC



**APPLICATION FOR  
UNITED STATES PATENT  
IN THE NAME OF**

**Jeffrey Skolnick, Mariusz Milik, and Andrzej Kolinski**

**of**

**The Scripps Research Institute**

**FOR**

**Prediction of Relative Binding Motifs of Biologically Active  
Peptides and Peptide Mimetics**

**John Land  
FISH & RICHARDSON  
4225 Executive Square, Suite 1400  
La Jolla, CA 92037  
(619) 678-5070 voice  
(619) 678-5099 fax**

Date of Deposit: 5/23/97

I hereby certify under 37 CFR 1.10 that this correspondence is being deposited with the United States Postal Service as "Express Mail Post Office To Addressee" with sufficient postage on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231.

Christopher Horne  
Christopher Horne

**DOCKET NO. 07300/034001**

**EXPRESS MAIL NO. EM12276061545**

70631 U.S. PTO

08/862192



05/23/97

PATENT APPLICATION SERIAL NO. \_\_\_\_\_

U.S. DEPARTMENT OF COMMERCE  
PATENT AND TRADEMARK OFFICE  
FEE RECORD SHEET

08862192

07/17/1997	BALEXAND	00000006	08862192
01	FC:101		770.00 OP
02	FC:102		240.00 OP
03	FC:103		88.00 OP
04	FC:104		260.00 OP

862192

-22-

~~08/862192~~

## ABSTRACT

A general neural network based method and system for identifying peptide binding motifs from limited experimental data. In particular, an artificial neural network (ANN) is trained with peptides with known sequence and function (*i.e.*, binding strength) identified from a phage display library. The ANN is then challenged with unknown peptides, and predicts relative binding motifs. Analysis of the unknown peptides validate the predictive capability of the ANN.

5

32278 LIT

08652192.1152397  
08652192.053397

## PREDICTION OF RELATIVE BINDING MOTIFS OF BIOLOGICALLY ACTIVE PEPTIDES AND PEPTIDE MIMETICS

### BACKGROUND OF THE INVENTION

#### 1. *Field of the Invention*

5 This invention relates to computer-assisted analysis of biological molecules, particularly of biologically active peptides and peptide mimetics.

#### 2. *Description of Related Art*

10 With the ever increasing plethora of biological information, the new branch of biological sciences called bioinformatics has become increasingly important. Bioinformatics seeks to translate the mass of protein (polypeptide) sequence information into knowledge of structure and more importantly, function.

One category of peptides where structure and function information would be useful are Class I major histocompatibility complex (MHC) molecules (in humans, the MHC is called HLA). MHC molecules are cell surface proteins that present bound peptides. These  
15 peptides are analyzed by immuno-surveillant cytotoxic T-cells (CTLs) to identify foreign or unhealthy cells for removal. Understanding this process is important, as it constitutes the primary immunological defense against viruses and perhaps tumor causing cells. It is also a major component responsible for transplant rejection. A. Townsend and H. Bodmer, *Annu. Rev. Immunol.* 7, 601 (1989); J.W. Yewdell and J.R. Binnink, *Adv.*  
20 *Immunol.* 52, 1 (1992). Since the affinity of the bound peptides largely determines the stability of the expressed class I molecules and their recognition by CTLs, it is crucial to determine the rules of peptide binding by class I molecules.

08362192-052397  
66250-26T2980

5

10

15

20

25

## SUMMARY OF THE INVENTION

The invention comprises a general neural network based method and system for identifying relative peptide binding motifs from limited experimental data. In particular, an artificial neural network (ANN) is trained with peptides with known sequence and function (*i.e.*, binding strength) identified from a phage display library. The ANN is then  
5 challenged with unknown peptides, and predicts relative binding motifs. Analysis of the unknown peptides validate the predictive capability of the ANN.

In one example, the training peptides bind to mouse MHC class I molecule H2-K<sup>b</sup>. Blind testing (*e.g.*, on chicken ovalbumin) correctly identified strongly binding peptides, and their relative binding strengths, in 5 of the 7 top scoring predictions from the test  
10 procedure. Upon validation analysis, the top scoring peptide was the known immunodominant peptide. Further, the second best binding peptide, since it lacked characteristic anchor residues, would have been missed using standard statistical approaches. The ability to predict antigens that bind MHC represents a significant advance in the  
15 development of vaccines and T-cell based therapeutics.

The details of the preferred embodiment of the present invention are set forth in the accompanying drawings and the description below. Once the details of the invention are known, numerous additional innovations and changes will become obvious to one skilled in the art.

088649-09397  
25129880



## 5-

5

5

FIGURE 3 is a graph showing a competition binding assay.

Like reference numbers and designations in the various drawings indicate like elements.

## 6

*Introduction*

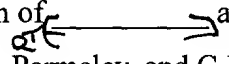
5 The invention will be described using an example of an artificial neural network (ANN) system used to predict relative binding motifs of peptides that bind to MHC class I molecules. However, the process is general and can be applied to any peptide system. An important aspect of the present invention is the inclusion of both experimental and theoretical aspects of the problem into one, coherent procedure. Preliminary results from the ANN analysis improved the interpretation of results from phage display experiments, and later experimental methods were used in blind tests of the ANN classification scheme.

Artificial neural networks can be used to recognize patterns and “signatures” in data streams. An ANN differs from other signal processing algorithms in that it does not assume any underlying model. Rather, an ANN “learns” to detect patterns by generating a model in response to input test data having known patterns, features, or other characteristics of interest in classifying the input data. An ANN can be trained relatively easy and repeatably. Because an ANN learns to detect patterns or correlations, ANNs are very flexible and adaptable to a wide variety of situations and conditions. This flexibility and adaptability gives artificial neural networks a significant advantage over other data classification techniques. For further information on the architecture and training of multi-layer perceptron (MLP) adaptive artificial neural networks, see “Progress in Supervised Neural Networks” by Don Hush and Bill Horne, published in *IEEE Signal Processing* (January 1993).

FIGURE 1 is a schematic view of the preferred peptide sequence coding scheme and the ANN architecture of the invention. Shown is a standard multi-layer perceptron ANN 1 trained by back-propagation (BP) of error. D. Rumelhart, J. McClelland and the PDP Research Group, "Parallel Distributed Processing", MIT Press, Cambridge (1986). The ANN 1 includes an input layer 2 comprising a plurality of input units 3, a hidden layer 4 comprising a plurality of hidden units 5, and an output layer 6 comprising a plurality of output units 7. In the preferred embodiment, the number of output units is two, denoted 7a and 7b. Each unit 3, 5, 7 is a processing element or "neuron", coupled by connections having adjustable numeric weights or connection strengths by which earlier layers influence later ones to determine the network output.

Prior to using the ANN 1 to classify actual input data, the parameters of the ANN 1 are adjusted by applying pre-characterized training data to the ANN 1. That is, training data is selected such that particular features are known to present or known to be absent. In the invention, such data comprises an appropriately coded set of input patterns (i.e., known peptide sequences having known binding affinities). See below for a discussion of the preferred coding.

#### Phage Display

In order to obtain training data for an ANN, a study was initiated with a peptide phage display binding analysis of the mouse MHC class I molecule K<sup>b</sup>. Soluble K<sup>b</sup> was purified from transfected Drosophila cells. Phage display analysis has been used previously to identify MHC class II molecule binding peptides. J. Hammer, B. Takacs and F. Sinigaglia, *J. Exp. Med.* 176, 1007 (1992). Phage display libraries were obtained from Dr. G.P. Smith of  and the analyses were performed essentially as described in the art (S.F. Parmeley, and G.P. Smith, *Gene* 73, 305 (1988); J.K. Scott and G. P. Smith, *Science* 249, 386 (1990); G.P. Smith personal communication). From the phage display, the sequences of 181 K<sup>b</sup> binding peptides and their relative binding affinities were obtained along with the sequences of 129 non-binding sequences.

08862192-052397  
08862192-052397

Ins e' >

## 8

15

190x

**TABLE 1A**

Clustering of amino acids according to their physico chemical features

No.	Feature	amino acid one-letter codes
0	hydrophobic	HWYFMLIVCAGTK
1	aliphatic	LIV
2	aromatic	FYWH
3	polar	TSNDEQURKHWY
4	charged	DERKH
5	positive	RKH
6	small	PVCAGTSND
7	tiny	AGS
8	glycine	G
9	proline	P

**TABLE 1B**

Feature based binary coding of amino acids

amino acid	feature based code
	0123456789
G	1000001110
A	1000001100
V	1100001000
L	1100000000
I	1100000000
S	0001001100
T	1001001000
D	0001101000
N	0001001000
K	1001110000
E	0001100000
Q	0001000000
R	0001110000
H	1011110000
F	1010000000
C	1000001000
W	1011000000
Y	1011000000
M	1000000000
P	0000001001

For example, in FIGURE 1, a peptide having the amino acid sequence of "SNPSFRPFA" is coded as a binary pattern beginning with the binary pattern for "S", and continuing

with the binary pattern for "N", *etc.* Of course, other mappings are possible, as well as other, fewer, and/or additional features.

### *ANN Training*

As indicated in FIGURE 1, the ANN 1 has two output nodes 7a, 7b. The output signal of the ANN 1 was defined as follows:

"00" (both nodes 7a, 7b off) denotes a non-binding sequence

"10" (first node 7a off, second node 7b on) denotes a weakly binding sequence

"11" (both nodes 7a, 7b on) denotes a strongly binding sequence.

The 181  $K^b$  binding peptides were divided into strong and weak binding classes, according to their respective experimentally measured binding constants. Additionally, the 129 peptides having no detectable affinity for  $K^b$  were used as negative examples. The entire 310 peptide data base was divided into training and testing sets. In this example, the testing set contained about 1/3 of the total number of peptides. A conjugate gradient procedure (T. Masters, "*Practical Neural Network Recipes in C++*", Acad. Press Inc. Boston (1993)) was used to determine the ANN weights, whose initial values were uniform pseudo-random numbers with a range of  $[-0.7, 0.7]$ . The network performance, defined as the mean square distance between the network output (*i.e.*, predicted binding strength) and experimentally observed value (*i.e.*, the known value of the binding strength), was measured as a function of the number of learning cycles or "epochs". One epoch occurs when the full set of training patterns is presented to the network.

FIGURE 2 is a graph showing performance of the experimental ANN 1 on the training and testing sets as a function of training time, measured by the number of epochs. As shown in FIGURE 2, while the error in the training set decreases monotonically with an increasing number of epochs, the testing set error reaches a minimum and then slowly grows as the ANN memorizes the training set, *i.e.*, as "over fitting" occurs. T. Masters, "*Practical Neural Network Recipes in C++*", Acad. Press Inc. Boston (1993). Thus, the ANN 1 weights were chosen where the error for the test set was approximately at a

0886249-0539  
2672980  
2672980

5

10

## 15

20

25

Peptide	Amino Acids	ANN	K <sub>D</sub> (moles/liter)	FACS Analysis % SIINFEKL
1	SIINFEKL	0.46	3.0E-9	100
2	SALAMVYL	0.44	7.1E-9	100
3	AEERYPIL	0.36	6.7E-5	42
4	NAIVFKGL	0.32	1.3E-8	76
5	KVVRFDKL	0.27	2.6E-8	94
6	RGDKLPGFG	0.26	5.5E-4	30
7	DVYSFSLA	0.24	7.0E-8	65
8	GTMSMLVL	0.23	1.2E-6	0
9	ASEKMKIL	0.22	5.5E-4	4
10	DHPFLFCI	0.20	4.7E-5	38
11	ENIFYCPI	0.19	9.4E-8	77
(VSV8)	RGYVYQGL	<i>no data</i>	4.1E-9	<i>not applicable</i>

11

12

10

15

**FACS Analysis.** Values from fluorescence activated cell sorter (FACS) analysis showing the relative amounts of K<sup>b</sup> on the surface of K<sup>b</sup> transfected drosophila cells following an 18-hour incubation with the indicated peptides. Cells were strained with the anti mouse MHC class 1 antibody Y3 followed by a fluoresceine conjugated second antibody. Median fluorescence values from separate experiments were normalized by subtracting the median fluorescence obtained in the absence of added peptides from each peptide sample and then expressing those values as the percent of the fluorescence obtained with SIINFEKL (which was examined in all experiments).

## 20

25



5

10

15

25

A list of 30 binding peptides were predicted along with scores for the predicted relative binding affinities. To evaluate these predictions, the 11 peptides at the top of the list were synthesized and their binding affinities determined experimentally. Our results demonstrate that the ANN 1 can make highly accurate predictions, some of which could not have been predicted manually using extant anchor position based binding rules. Five of the predicted seven best binders bound with good affinity ( $K_D < 10^{-7}$  nM). Most significantly, the top predicted peptide bound the strongest and is the known immunodominant epitope. Furthermore, despite the fact that the second best predicted peptide lacked internal anchor residues and thus would not have been included in the set of 20 manually predicted sequences, it was shown experimentally to bind with the second strongest affinity. This affinity is greater than four other predicted binding peptides in the top eleven scores, which do contain internal anchor residues.

Two peptides in the top 7 did not bind  $K^b$  with significant affinity; the question is why. One possibility is that binding to phage somehow does not accurately simulate peptide binding in all cases. Other possible reasons for these nonbinding sequences are that an insufficiently diverse combination of amino acids was present in the positive and negatively selected phage sequences or that the system of encoding amino acids for the ANN did not adequately distinguish the chemical and physical properties of all of the amino acids. These alternatives are presently being analyzed to improve accuracy of the invention. However, the success rate in the top seven predictions shows that the ANN approach works well.

In its present application, the ANN analysis should be able to predict class I binding peptides for an unlimited number of protein antigens. This may further the understanding of the class I molecular structure as it pertains to peptide binding and perhaps further elucidate how these binding interactions pertain to function. More generally, the inventive approach represents but a first application for identifying binding motifs from either peptide or even small molecule (*e.g.*, peptide mimetics) combinatorial libraries. One

strength of the invention is that it allows one to generalize and extract the latent information encoded in a random peptide library that has been screened for a particular property or functionality. The results of applying the ANN 1 of the invention may be used to design stronger binding sequences.

5     *Implementation*

10     The ANN 1 of the invention may be implemented in hardware or software, or a combination of both. However, preferably, the invention is implemented in computer programs executing on programmable computers each comprising at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. Program code is applied to input data to perform the functions described herein and generate output information. The output information is applied to one or more output devices, in known fashion.

15     Each program is preferably implemented in a high level procedural or object oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

20     Each such computer program is preferably stored on a storage media or device (e.g., ROM or magnetic diskette) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

25

08852192 052397  
462250 2672540

5

## CLAIMS

What is claimed is:

sub a<sup>2</sup> 1.

A method for identifying relative binding motifs of peptide-like molecules, comprising the steps of:

- (a) training an artificial neural network (ANN) with a set of training peptide-like molecules, each of known sequence and binding affinity;
- (b) applying to the ANN at least one peptide-like molecule, each of known sequence but unknown binding affinity;
- (c) analyzing each applied test peptide-like molecule using the ANN to predict a relative binding affinity for each test peptide-like molecule.

2. A method for identifying relative peptide binding motifs, comprising the steps of:

- (a) training an artificial neural network (ANN) with a set of training peptides, each of known binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
- (b) applying to the ANN at least one peptide, each of unknown binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
- (c) analyzing each applied test peptide using the ANN to predict a relative binding affinity for each test peptide.

08362192.052397  
66250.2612880

3. The method of claim 2, wherein the set of training peptides include peptides having a binding affinity for MHC class I molecules.
4. The method of claim 3, wherein the peptides included in the set of training peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
5. The method of claim 2, wherein the set of test peptides include peptides having a binding affinity for MHC class I molecules.
6. The method of claim 5, wherein the peptides included in the set of test peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
7. The method of claims 1 or 2, wherein the ANN comprises a multi-layer perceptron ANN trained by back-propagation of error.

20250726 10:30:00  
20250726 10:30:00

8. A system for identifying relative binding motifs for peptide-like molecules, comprising:
- (a) means for training an artificial neural network (ANN) with a set of training peptide-like molecules, each of known sequence and binding affinity;
  - (b) means for applying to the ANN at least one test peptide-like molecule, each of known sequence but unknown binding affinity;
  - (c) means for analyzing each applied test peptide-like molecule using the ANN to predict a relative binding affinity for each test peptide-like molecule.
9. A system for identifying relative peptide binding motifs, comprising:
- (a) means for training an artificial neural network (ANN) with a set of training peptides, each of known binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
  - (b) means for applying to the ANN at least one test peptide, each of unknown binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
  - (c) means for analyzing each applied test peptide using the ANN to predict a relative binding affinity for each test peptide.

10. The system of claim 9, wherein the set of training peptides include peptides having a binding affinity for MHC class I molecules.
11. The system of claim 10, wherein the peptides included in the set of training peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
12. The system of claim 9, wherein the set of test peptides include peptides having a binding affinity for MHC class I molecules.
13. The system of claim 12, wherein the peptides included in the set of test peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
14. The system of claims 8 or 9, wherein the ANN comprises a multi-layer perceptron ANN trained by back-propagation of error.

46E250" 26T29880

46E250" 26T29880



15. A computer program, residing on a computer-readable medium, for identifying relative binding motifs for peptide-like molecules, comprising instructions for causing a computer to:
- (a) train an artificial neural network (ANN) with a set of training peptide-like molecules, each of known sequence and binding affinity;
  - (b) apply to the ANN at least one test peptide-like molecule, each of known sequence but unknown binding affinity;
  - (c) analyze each applied test peptide-like molecule using the ANN to predict a relative binding affinity for each test peptide-like molecule.
16. A computer program, residing on a computer-readable medium, for identifying relative peptide binding motifs, comprising instructions for causing a computer to:
- (a) train an artificial neural network (ANN) with a set of training peptides, each of known binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
  - (b) apply to the ANN at least one test peptide, each of unknown binding affinity, each peptide comprising a sequence of amino acids, each amino acid being binary coded as having or lacking specific features generally characteristic of amino acids;
  - (c) analyze each applied test peptide using the ANN to predict a relative binding affinity for each test peptide.

5

5

10

17. The computer program of claim 16, wherein the set of training peptides having a binding affinity for MHC class I molecules.
18. The computer program of claim 17, wherein the peptides included in the set of training peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
19. The computer program of claim 16, wherein the set of test peptides include peptides having a binding affinity for MHC class I molecules.
20. The computer program of claim 19, wherein the peptides included in the set of test peptides have a binding affinity for mouse MHC class I K<sup>b</sup>.
21. The computer program of claims 15 or 16, wherein the ANN comprises a multi-layer perceptron ANN trained by back-propagation of error.